Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

CHESSBOARD DISTRIBUTIONS

Soumyadip Ghosh
Shane G. Henderson

School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853, U.S.A.

ABSTRACT

We review *chessboard distributions* for modeling partially specified finite-dimensional random vectors. Chessboard distributions can match a given set of marginals, a given covariance structure, and various other constraints on the distribution of a random vector. It is necessary to solve a potentially large linear program to set up a chessboard distribution, but random vectors can then be rapidly generated.

1 INTRODUCTION

In the construction of stochastic models, it is often the case that dependence among the primitive inputs is essential. Ignoring this dependence and using independent inputs can result in a model that might not be valid for the system being modelled:

1. Hodgson et al. (2000) observe that correlations between various factors play an important role in determining the statistical properties of *real* job shop scheduling problems, which are fundamentally different in nature from those of problems that are generated artificially.
2. Cost estimation for new projects are nowadays often accompanied by an analysis of the risk of cost overruns (Lurie and Goldberg 1998). The analysis begins by assessing the uncertainty in the individual elements and then these factors are aggregated into a distribution of the total cost. Dependence among the various elements can be very important to this aggregation.

There is hence a growing realization of the need to consider the dependence among input random variables when generating samples for a simulation of a stochastic model. This translates into the need for a method that can efficiently generate samples of correlated random variables. In our study we look at the case where the correlated primitive inputs of a model are finite in number and hence can be characterized jointly as a random vector.

For the case of a finite-dimensional random vector, its full joint distribution gives complete information about the dependencies among its components. Hence the ideal approach would be to try to model and generate from the complete joint distribution of the random vector. Indeed several such methods have been suggested, but almost all of them suffer from some serious drawbacks. Some methods need an enormous amount of information for modelling the joint distribution and are hence efficiently implemented only for random vectors of low dimensionality. These methods are also usually very specific in nature and cannot be easily adapted to handle any arbitrary joint distribution. Hence these methods can become impractical for a model of even moderate complexity.

A practical alternative is to specify just the marginal distributions of the components of the random vector along with some reasonable measure of dependence, like the correlation or covariance matrix, and require that a joint distribution be constructed that satisfies the given criteria. This joint distribution is not necessarily the unique joint distribution that fully describes the situation, but might be a reasonable (relative to the application at hand) approximation. The covariance measure could be Spearman's rank covariance, Pearson's product-moment covariance, Kendall's τ, or any other convenient covariance measure.

Hill and Reilly (1994, 2000) describe a method for constructing joint distributions as probability mixtures or convex compositions of extremal distributions and the independent joint distribution. Extremal distributions are those that induce the extreme possible correlations between its components. Whitt (1976) gave the main analytical results for the bivariate case. While this method is very effective for random vectors of low dimensions ($d <= 3$ say), the computational requirements quickly become expensive for higher dimensional random vectors.

Meeuwissen and Bedford (1997) study minimally informative bivariate distributions with uniform marginals and

a specified rank correlation. These are distributions that have the maximum entropy, and hence the minimum information, possible with respect to the independent joint distribution subject to the constraint of having the right rank correlation. Minimally informative distributions are used in constructing tree-dependent and vine-dependent random vectors. Tree-dependent random vectors (Meeuwissen and Cooke 1994) are constructed from an undirected acyclic graph and a set of bivariate distributions for those pairs of components which are connected in the graph. This method can generate samples on the fly, and is very easily implemented. But the input rank correlations this method can match is limited in number to a maximum of $d-1$ for a d-dimensional random vector (Cooke 1997). Cooke (1997) introduces the vine-dependent random vectors as a generalization of the tree-dependent random vectors to remedy this limiting disadvantage while maintaining its advantage of extremely fast generation.

Cario and Nelson (1997) described a method for generating random vectors with prescribed Pearson product-moment covariance matrix, which they called the "NORmal To Anything" (NORTA) method. Variations and specializations of the NORTA method are developed in Mardia (1970), Li and Hammond (1975) and Iman and Conover (1982). The NORTA method first starts with a correlated joint normal random vector and transforms each individual component through an appropriate inversion function to a corresponding random variable with the desired marginal distribution. The correlation matrix of the random vector that results from the transformation is different from the correlation matrix of the joint normal vector due to the (generally nonlinear) transformation. Hence, the correlation matrix for the joint normal vector has to be chosen such that the correlations that result from the transformations are exactly those desired. Cario and Nelson (1997) outline a numerical method for determining the right correlation matrix for the joint normal random vector.

The NORTA method capitalizes on the fact that multivariate correlated normal vectors are easily generated. It is one of the most efficient general purpose methods available in practice to generate random vectors. Some of its attractive features include its generality and ease of implementation. The NORTA procedure, when it works, is often the method of choice for generating random vectors with arbitrary marginals and any given feasible covariance matrix.

One can, however, experience a difficulty while using the approaches mentioned above. We say that a covariance matrix is *feasible* for a given set of marginal distributions if there exists a random vector (that is, a joint distribution) with the prescribed marginals and covariance matrix.

A matrix must necessarily be positive semidefinite to be a feasible covariance matrix, but this condition is not, in general, sufficient. Consider a random vector that consists of an exponential random variable with mean 1 and a uniform (0,1] random variable. The maximum (minimum) correlation that can be induced between the two is strictly less (greater) than +1 (-1). This is because if two random variables have a correlation of ± 1, then one is a linear function of the other, which clearly cannot hold in this case. This is also why two exponential random variables cannot achieve a negative correlation of -1. In general, the range of correlations that can be achieved between two random variables is a strict subinterval of the interval [-1,1]. Hence determining the feasibility or not of a given matrix for a specified set of marginals is not trivial.

Ghosh and Henderson (2001) developed a computational technique that was then used to show that there are sets of marginals with feasible covariance matrix that cannot be matched using either the extremal distributions technique or the NORTA method. (It is not yet clear whether the vine-dependent random vector technique can be used to model *all* feasible covariance matrices.)

The computational technique involves solving a linear program to try and construct a chessboard distribution, as defined in Section 2, that matches the desired marginals and covariance matrix. We review this technique and its properties in the remainder of this paper.

Chessboard distributions are perhaps closest in nature to the "piecewise-uniform copulae" that Mackenzie (1994) develops. Mackenzie (1994) assumes that a covariance matrix is feasible for uniform marginal distributions, and then proceeds to try to identify one with maximum entropy. In contrast, we do not assume feasibility but develop this family of distributions as a tool to investigate the feasibility or not of given covariance matrices.

The next section reviews the construction of a chessboard distribution and Section 3 shows how to match a given measure of dependence. Section 4 summarizes the key theoretical results that we have been able to derive regarding the properties of these distributions. Section 5 deals with an extension where we try to match other information available regarding the distribution of the random vector. A consequence of the computational matching procedure is that when a random vector exists, the procedure returns an explicit construction of a chessboard distribution for that random vector, which, it turns out, can be rapidly generated from. Section 6 deals with the issues involved in generating from a constructed distribution. Section 7 details topics of ongoing research.

2 CONSTRUCTING A CHESSBOARD DISTRIBUTION

Suppose that we wish to generate a random vector with 3 components. The extension to the general d component case with $d \geq 2$ is straightforward. Let F_i denote the desired marginal distribution function of the ith component, for

$i = 1, 2, 3$. In this paper we assume that the distributions F_i have densities f_i (with respect to Lebesgue measure) for all i. However, much of what follows can be generalized.

We construct a random vector $X = (X_1, X_2, X_3)$ where the joint distribution of X has a special structure. We divide $\Re^3$ into a large grid of rectangular regions (cells) with sides parallel to the coordinate axes. Within each cell, the components of X are conditionally independent with marginal distributions given by the F_is restricted to the cell. Let $n \geq 1$ denote a parameter that is used to determine the cells. We divide each coordinate direction into n regions.

To be precise, for $i = 1, 2, 3$ let

$$-\infty \leq y_{i0} < y_{i1} < \ldots < y_{in} \leq \infty$$

denote breakpoints on the ith axis such that for $j = 1, 2, \ldots, n$

$$F_i(y_{ij}) - F_i(y_{i,j-1}) = 1/n.$$

Note that we explicitly allow $y_{i0} = -\infty$ and $y_{in} = \infty$ to capture distributions with unbounded support. The breakpoints divide the ith coordinate axis into n intervals of equal probability with respect to the marginal distribution function F_i for $i = 1, 2, 3$.

For $1 \leq j_1, j_2, j_3 \leq n$ define the cell $C(j_1, j_2, j_3)$ to be the (j_1, j_2, j_3)th rectangular region

$$\{x = (x_1, x_2, x_3) : y_{i,j_i-1} < x_i \leq y_{i,j_i}\, i = 1, 2, 3\} \cap \Re^3.$$

Define

$$q(j_1, j_2, j_3) = P(X \in C(j_1, j_2, j_3))$$

to be the probability that the constructed random vector appears in the (j_1, j_2, j_3)th cell. To be consistent with the given marginals, the $q(j_1, j_2, j_3)$ values must satisfy the constraints

$$\begin{aligned}
\sum_{j_2=1}^{n}\sum_{j_3=1}^{n} q(j_1, j_2, j_3) &= 1/n \qquad j_1 = 1, \ldots, n \\
\sum_{j_1=1}^{n}\sum_{j_3=1}^{n} q(j_1, j_2, j_3) &= 1/n \qquad j_2 = 1, \ldots, n \qquad (1) \\
\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} q(j_1, j_2, j_3) &= 1/n \qquad j_3 = 1, \ldots, n \\
q(j_1, j_2, j_3) &\geq 0 \qquad 1 \leq j_1, j_2, j_3 \leq n.
\end{aligned}$$

The density $f(x)$ of X evaluated at $x \in C(j_1, j_2, j_3)$ is then given by

$$q(j_1, j_2, j_3)\frac{f_1(x_1)}{1/n}\frac{f_2(x_2)}{1/n}\frac{f_3(x_3)}{1/n}, \qquad (2)$$

so that conditional on $X \in C(j_1, j_2, j_3)$, the components of X are independent, and are distributed according to the desired marginals f_1, f_2, f_3 restricted to the cell $C(j_1, j_2, j_3)$. But does X then have the desired marginals?

Theorem 1. *If q satisfies the constraints (1), and X is constructed with density f as given in (2), then X has the desired marginals.*

A proof of this result may be found in Ghosh and Henderson (2001). We repeat the proof because it is useful in understanding the nature of chessboard distributions.

Proof: Clearly, f is nonnegative and integrates to 1. We need only show that the marginals of f are as specified to complete the proof. Let the marginal density function of X_1 be denoted by $g_1(\cdot)$. Then we have to prove that g_1 is equal to f_1. For any $x \in (y_{1,j_1-1}, y_{1,j_1})$, we have that

$$\begin{aligned}
&g_1(x)dx \\
&= \sum_{j_2,j_3=1}^{n} P(X_1 \in [x, x+dx) | X \in C(j_1, j_2, j_3)) \times \\
&\quad P(X \in C(j_1, j_2, j_3)) \\
&= \sum_{j_2,j_3=1}^{n} P(X_1 \in [x, x+dx) | X_1 \in (y_{1,j_1-1}, y_{1,j_1}]) \times \\
&\quad q(j_1, j_2, j_3) \\
&= \sum_{j_2,j_3=1}^{n} \frac{f_1(x)\,dx}{1/n} q(j_1, j_2, j_3) \\
&= nf_1(x)\,dx \sum_{j_2,j_3=1}^{n} q(j_1, j_2, j_3) = f_1(x)\,dx
\end{aligned}$$

The first equation follows by conditioning on the cell in which the random vector lies, and the second by the conditional independence of X_1, X_2 and X_3 given that X lies in $C(j_1, j_2, j_3)$. The final equation follows from (1). A similar result holds for the marginals of X_2 and X_3, and so the joint density f has the desired marginals. □

3 MATCHING COVARIANCE

There are several measures of correlation/covariance between random variables. Two of the most popular are Pearson's product moment covariance and Spearman's rank correlation. The Pearson product-moment covariance be-

tween two random variables V and W is given by

$$\text{cov}(V, W) = E(VW) - EVEW$$

and is well-defined if $E(V^2 + W^2) < \infty$. The Spearman rank covariance between V and W is given by

$$\begin{aligned} \text{rcov}(V, W) &= \text{cov}(F(V), G(W)) \\ &= E[F(V)G(W)] - E[F(V)]E[G(W)], \end{aligned}$$

where F and G are the distribution functions of V and W respectively. In contrast to product-moment covariance, the rank covariance is always well-defined and finite since $F(V)$ and $G(W)$ are bounded random variables. Indeed, if F is continuous, then $F(V)$ has a uniform distribution on the interval $(0, 1)$. Another attractive property of rank covariance that is not shared by product-moment covariance is that rank covariance is preserved under strictly increasing transformations, i.e., if h is a strictly increasing function, then $\text{rcov}(h(V), W) = \text{rcov}(V, W)$.

We give methods for matching both rank and product-moment covariance. In both cases we restrict attention to the case where the marginal distributions all have densities. First we consider rank covariance.

3.1 RANK COVARIANCE

Let X be a random vector of the form introduced in the previous section. Let Σ_X denote the *actual* rank covariance matrix of X, and let Σ denote the *desired* rank covariance matrix of X. In order to match Σ_X to Σ, it suffices to look at the elements above the diagonal, since covariance matrices are symmetric, and the diagonal elements of Σ_X are $1/12$ (recall that $F_i(X_i)$ is uniformly distributed on $(0, 1)$, since F has a density and is therefore continuous). Our approach is to construct a random vector as given in the previous section that minimizes the distance $r(\Sigma, \Sigma_X)$ between Σ and Σ_X, where

$$r(\Sigma, \Sigma_X) = \sum_{1 \le i < j \le 3} |\Sigma(i, j) - \Sigma_X(i, j)| .$$

In other words, we wish to

$$\min r(\Sigma, \Sigma_X), \tag{3}$$

subject to the constraints (1).

The rank covariance $\Sigma_X(1, 2)$ between X_1 and X_2 is given by

$$\begin{aligned} & E[F_1(X_1)F_2(X_2)] - E[F_1(X_1)]E[F_2(X_2)] \\ &= E[F_1(X_1)F_2(X_2)] - \frac{1}{4} \qquad (4) \\ &= \sum_{j_1, j_2, j_3} E[F_1(X_1)F_2(X_2)|X \in C(j_1, j_2, j_3)] \times \\ &\quad q(j_1, j_2, j_3) - \frac{1}{4} \\ &= \sum_{j_1, j_2, j_3} \{ E[F_1(X_1)|X \in C(j_1, j_2, j_3)] \times \qquad (5) \\ &\quad E[F_2(X_2)|X \in C(j_1, j_2, j_3)] \} q(j_1, j_2, j_3) - \frac{1}{4} \\ &= \sum_{j_1, j_2, j_3} \frac{(2j_1 - 1)}{2n} \frac{(2j_2 - 1)}{2n} q(j_1, j_2, j_3) - \frac{1}{4}. \qquad (6) \end{aligned}$$

Equation (4) follows since $F_i(X_i)$ have uniform distributions on $(0, 1)$ for all i. The equality (5) holds because conditional on X lying in a specified cell, the components of X are independent. The final equality (6) holds because conditional on X lying in cell $C(j_1, j_2, j_3)$, $F_i(X_i)$ is uniformly distributed on the interval

$$\left(\frac{j_i - 1}{n}, \frac{j_i}{n} \right),$$

for $i = 1, 2, 3$.

Equation (6) gives the rank covariance between X_1 and X_2 as a linear function of $q(\cdot, \cdot, \cdot)$. We can do the same for X_2 and X_3, and X_1 and X_3. Using standard linear programming techniques, we can then reexpress (3) to remove the nonlinear absolute values. In particular, we can use the linear program

$$\begin{aligned} \min \quad & \textstyle\sum_{1 \le i < j \le 3} z_{ij}^+ + z_{ij}^- \\ & z_{ij}^+ - z_{ij}^- = \Sigma(i, j) - \Sigma_X(i, j) \quad i < j \\ & z_{ij}^+, z_{ij}^- \ge 0 \quad i < j \end{aligned}$$

together with the constraints (1) to identify the distribution of X.

For a d-dimensional random vector, this linear program has $n^3 + d(d-1)$ variables and $d(d-1)/2 + dn$ constraints. In the 3-dimensional case, the variables are the $q(j_1, j_2, j_3)$ $(1 \le j_1, j_2, j_3 \le n)$ which appear implicitly in the term $\Sigma_X(i, j)$, and the z_{ij}^+, z_{ij}^- $(1 \le i < j \le 3)$.

We can tighten the feasible region of the linear program through the addition of bounds on the z_{ij}^+ and z_{ij}^- variables. These bounds are obtained by first assuming the existence of a random vector Z with the desired marginals and rank covariance matrix Σ. We then convert Z into another random

vector $\tilde{Z}$ that gives a feasible solution to the above LP. This is done by setting $q(j_1, j_2, j_3) = P(Z \in C(j_1, j_2, j_3))$ for all j_1, j_2, j_3, and then creating the density of $\tilde{Z}$ from $q(\cdot, \cdot, \cdot)$ as in the previous section. One can bound the change in the rank covariance matrix when transforming from Z to $\tilde{Z}$ in this fashion, and these bounds translate into bounds on the z_{ij}^+ and z_{ij}^- variables. See Ghosh and Henderson (2001) for details.

3.2 PRODUCT-MOMENT COVARIANCE

Now suppose that the marginals of X have finite second moments, and we want the *actual* product-moment covariance matrix Σ_X to match the *desired* product-moment covariance matrix Σ. The diagonal elements of the covariance matrices Σ and Σ_X are determined by the marginal distributions, and so our objective is again to

$$\min r(\Sigma, \Sigma_X),$$

subject to the constraints (1).

Using similar reasoning to that used for computing the rank covariance, we find that for $i \neq k$,

$$\Sigma_X(i,k) = \sum_{j_1, j_2, j_3} \gamma_i(j_i)\gamma_k(j_k)q(j_1, j_2, j_3) - EX_i EX_k, \tag{7}$$

where, for $1 \leq i \leq 3$ and $1 \leq m \leq n$,

$$\gamma_i(m) = E[X_i | X_i \in (y_{i,m-1}, y_{im}]]$$

is the conditional mean of X_i given that it lies in the mth subinterval.

Thus, we have again expressed $\Sigma_X(i,k)$ as a linear function of q. To match the desired covariance matrix Σ, we solve the linear program

$$\begin{aligned} \min \quad & \textstyle\sum_{1 \leq i < j \leq 3} z_{ij}^+ + z_{ij}^- \\ & z_{ij}^+ - z_{ij}^- = \Sigma(i,j) - \Sigma_X(i,j) \quad i < j \\ & z_{ij}^+, z_{ij}^- \geq 0 \quad i < j \end{aligned}$$

together with the constraints (1). Here, we use the expression (7) for Σ_X in place of (6).

In the case where all of the marginal distributions have bounded support, we can tighten the feasible region through the addition of explicit bounds on the z_{ij}^+ and z_{ij}^- variables (Ghosh and Henderson 2001). These additional bounds are important, because they enable one to provide a "feasibiliy check" for a given covariance matrix; see Section 4. The bounds in Ghosh and Henderson (2001) require that the maximum side-length of a cell converges to 0 as $n \to \infty$. This may not be true with our existing cell layout if one of the marginal distribution functions has a "flat patch", i.e., there is some i and $x < y$ such that $F_i(x) = F_i(y)$. However, if we redesign the cell layout so that the maximum side-length of a cell converges to 0 as $n \to \infty$, which is easily done, then we can add explicit bounds.

When some of the marginals have unbounded support, we do not have explicit bounds on these variables.

4 PROPERTIES

We have now introduced linear programs for constructing chessboard distributions with given marginals and covariance matrix. But how effective are these methods? In this section we summarize key results from Ghosh and Henderson (2001) on this topic without proof. In giving these results, we allow the random vector to have arbitrary, but finite, dimension $d > 1$. We restrict attention to marginal distributions with densities and bounded support, but we believe that most of the results can be established under weaker conditions.

Definition 1. *We say that a product-moment covariance matrix* Σ *is* product-moment-feasible *for a given set of marginal distributions if a random vector can be constructed with the given marginals and product-moment covariance matrix.*

We can similarly define *rank-feasible* covariance matrices for a given set of marginals.

Theorem 2. *Suppose that all of the marginal distributions have densities and bounded support. Then a covariance matrix is product-moment-infeasible (rank-infeasible) for the marginals if, and only if, the chessboard LP constructed earlier to match product-moment covariances (rank covariances) is infeasible for some* $n \geq 1$.

This result establishes that if one of the LPs is infeasible *for any discretization level* n, then the proposed covariance matrix is infeasible. Furthermore, the theorem establishes that if a covariance matrix is infeasible, then one will eventually discover this by solving an LP with n sufficiently large.

To our knowledge, this is the first example of a tight characterization of infeasible covariance matrices for random vectors of dimension $d \geq 3$.

Of course, we are more interested in a positive result. Given the sharp characterization in Theorem 2, it would be nice if chessboard distributions could *exactly* match any arbitrary feasible covariance matrix. Unfortunately, this is not the case, as the following example shows.

Example 1. *Suppose that* $Z_1 = Z_2$ *is uniformly distributed on* $(0, 1)$, *so that* $cov(Z_1, Z_2) = var(Z_1) = 1/12$. *For given n, the covariance between* X_1 *and* X_2 *is maximized by concentrating all mass on the cells* (i, i), *and so* $q(i, i) =$

n^{-1} *for* $1 \leq i \leq n$. *In that case, we have that*

$$cov(X_1, X_2) = \frac{1}{12} - \frac{1}{12n^2}$$

Therefore, $cov(X_1, X_2) < 1/12$ *for all finite* n.

This example shows that chessboard distributions cannot exactly match all feasible covariance matrices. Notice though, that the error in the covariance matrix can be made arbitrarily small. In fact it is possible to show, for certain marginals, that chessboard distributions can arbitrarily closely approximate any feasible covariance matrix.

Theorem 3. *Suppose that the marginal distributions have densities and bounded support, and that* Σ_{pm} *(*Σ_{r}*) is product-moment (rank) feasible. Then for all* $\epsilon > 0$*, there exists a chessboard distribution with product-moment (rank) covariance matrix* Λ_{pm} *(*Λ_{r}*) with the property that* $r(\Sigma_{\text{pm}}, \Lambda_{\text{pm}}) < \epsilon$ *(*$r(\Sigma_{\text{r}}, \Lambda_{\text{r}}) < \epsilon$*).*

Not only can chessboard distributions closely approximate any feasible covariance matrix, but they can *exactly* match "almost all" feasible covariance matrices. To formulate and state this result precisely, we need some more terminology and a definition.

Suppose that the marginal distributions $F_1, \ldots, F_d$ are fixed. We can, with an abuse of notation, view a $d \times d$ covariance matrix as an element of $d(d-1)/2$ dimensional space. This follows because there are $d(d-1)/2$ elements above the diagonal, the matrix is symmetric, and the diagonals are determined by the marginal distributions. Let Ω_{pm} (Ω_{r}) denote the set of feasible product-moment (rank) covariance matrices. We view these sets as subsets of $d(d-1)/2$ dimensional space. Ghosh and Henderson (2001) prove the following two results.

Proposition 4. *If the marginal distributions have densities and bounded support, then the sets* Ω_{pm} *and* Ω_{r} *are nonempty, convex, closed and full-dimensional.*

Let A° denote the interior of the set A.

Theorem 5. *Suppose that the marginal distributions have densities and bounded support. Then there is a chessboard distribution with product-moment (rank) covariance matrix* Σ_{pm} *(*Σ_{r}*) if, and only if,* $\Sigma_{\text{pm}} \in \Omega_{\text{pm}}^\circ$ *(*$\Sigma_{\text{r}} \in \Omega_{\text{r}}^\circ$*).*

In summary then, under certain assumptions on the marginal distributions, chessboard distributions

- can arbitrarily closely approximate any feasible covariance matrix,
- can exactly match any feasible covariance matrix in the interior of the set of feasible covariance matrices, but
- cannot exactly match any covariance matrix on the boundary of the set of feasible covariance matrices.

We have stated all of these results under the assumption that the marginal distributions have densities and bounded support. We believe that the results apply more generally, and are working on extending these results to more general marginals.

5 OTHER CONSTRAINTS

In the previous sections we developed a method for constructing a partially specified random vector and stated some of its theoretical properties. In particular, we looked at random vectors with prescribed marginals and covariance matrix. But one may have other forms of information about the random vector. Clemen, Fischer and Winkler (2000) surveyed and presented a variety of measures chosen primarily for their potential use in eliciting information from experts. Among those presented were correlations and the measures presented below.

Those measures that, for chessboard distributions, can be expressed as linear functions of the $q(j_1, j_2, j_3)$s can be incorporated as requirements into our LPs either as constraints or as part of the objective function. Therefore, for each measure, we consider whether it can be expressed in linear form or not. For simplicity, we assume that all of the marginals have densities.

1. (Joint probabilities.) For fixed $i \neq j, b_i$ and b_j, let

   $$\text{JP}_{ij}(b_i, b_j) = P(X_i \leq b_i, X_j \leq b_j)$$

   be the joint probability that X_i is at most b_i and X_j is at most b_j. Then $\text{JP}_{ij}(b_i, b_j)$ can be represented as a linear function of the variables $q(\cdot, \cdot, \cdot)$ introduced in Section 2. This is particularly straightforward if b_i and b_j coincide with breakpoints between cells, but also holds even if this is not the case. Let $i = 1$ and $j = 2$. Then, for any $x = (x_1, x_2, x_3)$, assuming that the points $(b_1, b_2, \cdot)$ belong to the

cells $C(J_1, J_2, \cdot)$, we can write

$$\begin{aligned}
&\mathrm{JP}_{12}(b_1, b_2) \\
&= \int_{-\infty}^{b_1}\int_{-\infty}^{b_2}\int_{-\infty}^{+\infty} f(x)\,dx \\
&= \sum_{j_1=1}^{J_1-1}\sum_{j_2=1}^{J_2-1}\sum_{j_3=1}^{n} q(j_1, j_2, j_3) \\
&\quad + nA_1(b_1)\sum_{j_2=1}^{J_2-1}\sum_{j_3=1}^{n} q(J_1, j_2, j_3) \\
&\quad + nA_2(b_2)\sum_{j_1=1}^{J_1-1}\sum_{j_3=1}^{n} q(j_1, J_2, j_3) \\
&\quad + n^2A_1(b_1)A_2(b_2)\sum_{j_3=1}^{n} q(J_1, J_2, j_3)
\end{aligned}$$

where, for $i = 1, 2, 3$,

$$A_i(b_i) = \int_{y_{i,J_i-1}}^{b_i} f_i(x_i)\,dx_i.$$

Clearly, this expression is linear in the variables $q(\cdot, \cdot, \cdot)$.

2. (Conditional fractiles.) For fixed $i \neq j$ and $0 < p < 1$, let

$$\mathrm{CF}_{ij}(p) = E[F_{i|j}(X_i)|F_j(X_j) = p]$$

be the conditional expectation of a percentile of X_i given that $F_j(X_j) = p$. Let $i = 1$ and $j = 2$. Let J_2 be such that $F_j^{-1}(p) \in \{y_{2,J_2-1}, y_{2,J_2}\}$. Then it is possible to show that

$$\begin{aligned}
\mathrm{CF}_{12}(p) &= E[F_{1|2}(X_1)|F_2(X_2) = p] \\
&= \frac{1}{2}\sum_{j_1=1}^{n}\sum_{j_3=1}^{n} q(j_1, J_2, j_3)\,(2j_1 - 1)
\end{aligned}$$

Thus, $\mathrm{CF}_{12}(p)$ is also a linear function of the variables $q(\cdot, \cdot, \cdot)$.

3. (Concordance probabilities). Let $i \neq j$ be fixed, and let X and Y be independent, identically distributed random vectors. Let

$$\mathrm{CNC}_{ij} = P(X_i > Y_i|X_j > Y_j).$$

This performance measure can be expressed as a quadratic function of the variables $q(\cdot, \cdot, \cdot)$, but not as a linear function. Therefore, to incorporate this measure into the chessboard distribution calculation, one would have to solve a quadratic program.

Extending the theoretical results of the previous section to incorporate LPs with these additional constraints is the subject of ongoing research.

6 GENERATION

How should one go about generating random vectors with a chessboard distribution? Many methods are possible. The methods vary in terms of their time and storage requirements for setup, and for generating random vectors once the setup is complete. In this section we sketch an approach that requires a moderate amount of time and storage for setup, but once the setup is complete requires very little time to generate random vectors.

Let d denote the dimension of the random vector X with marginal distribution functions $F_1, \ldots, F_d$. Suppose that $q(\cdot, \ldots, \cdot)$ is the solution to one of the linear programs discussed earlier. We propose the following 2-step procedure for generating chessboard random vectors with these characteristics.

1. Generate the indices $(j_1, \ldots, j_d)$ of the cell containing X from the probabilities $q(\cdot, \ldots, \cdot)$.
2. Generate X from its conditional distribution given that it lies in the cell $C(j_1, \ldots, j_d)$.

The first step can be performed efficiently using, for example, the alias method. The alias method, developed by Walker (1977) and discussed in detail in Law and Kelton (2000), can generate the appropriate cell in constant time, and requires $O(m)$ storage and $O(m)$ setup time, where m is the number of positive $q(j_1, j_2, \ldots, j_d)$ values. If $q(\cdot, \ldots, \cdot)$ is an extreme-point solution to one of the linear programs developed earlier, then there are on the order of nd strictly positive cell probabilities. This follows from a standard result in linear programming that any extreme point solution to a system of m linear equalities in nonnegative variables has at most m strictly positive values. The exact number of positive values depends on the number of equality constraints in the LP and the degree to which the extreme-point solution is degenerate. (A degenerate extreme-point solution is one with less than m strictly positive values.)

The fact that $m = O(nd)$ is relatively small can be viewed as an advantage with respect to variate generation since it reduces the setup time required to implement the alias method. However, it can also be viewed as a disadvantage in terms of modeling power. For a given dimension d and discretization level n there are n^d cells. Of these, $O(nd)$ receive strictly positive probabilities $q(j_1, \ldots, j_d)$. So as the dimension d increases, the fraction of cells receiving positive probabilities is vanishingly small. This means that

the set of values that the random vector X can assume is somewhat limited.

Mackenzie (1994) prevents this situation from arising by maximizing the entropy of the discrete distribution $q(\cdot, \ldots, \cdot)$. In this case, all of the cells receive positive probability. However, the problem of maximizing the entropy of q subject to linear constraints is a convex optimization problem that is more difficult to solve than the LPs discussed in this paper. The degree to which this "sparsity" issue is, indeed, an issue is the subject of current research, as are methods to deal with it.

Let $C(j_1, \ldots, j_d)$ denote the cell chosen in Step 1 above. Conditional on X lying in this cell, the components $X_1, \ldots, X_d$ of X are conditionally independent. Thus, in Step 2, we can independently generate each component from its respective conditional (marginal) distribution. Any standard univariate generation method can be specialized to this problem, including inversion and acceptance-rejection.

7 FUTURE RESEARCH

In this paper we have described some of the key ideas and results on chessboard distributions. Several questions remain the subject of ongoing research.

1. To what degree can one relax the assumptions on the marginal distributions and still prove results of the form presented here?
2. Is it possible to extend the notion of a chessboard distribution to enable the generation of "boundary" covariance matrices?
3. What other aspects of joint distributions might be specified by a user, and can chessboard distributions be selected to capture this information?
4. The linear programs that we need to solve to construct a chessboard distribution could be very large. Can they be solved more efficiently than through a general-purpose linear programming code?
5. Is the "sparsity" issue discussed in Section 6 a problem, and if so, are there elegant and appropriate methods for resolving it?

ACKNOWLEDGMENTS

This research was partially supported by National Science Foundation Grant Number DMI 9984717.

REFERENCES

Cario, M. C. and B. L. Nelson. 1997. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Clemen, R. T., G. W. Fischer, and R. L. Winkler. 2000. Assessing dependence: some experimental results. *Management Science* 46: 1100–1115.

Cooke, R. M. 1997. Markov and entropy properties of tree- and vine-dependent variables. *Proceedings of the ASA Section on Bayesian Statistical Science.* Alexandria, VA.

Ghosh, S., and S. G. Henderson. 2001. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*. Submitted.

Hill, R. R., and C. H. Reilly. 1994. Composition for multivariate random vectors. In *Proceedings of the 1994 Winter Simulation Conference*, J. D. Tew, S. Manivannan, D. A. Sadowsky, A. F. Seila, eds. IEEE, Piscataway New Jersey, 332 – 339.

Hill, R. R., and C. H. Reilly. 2000. The effects of coefficient correlation structure in two-dimensional knapsack problems on solution procedure performance. *Management Science*, 46: 302–317.

Hodgson, T. J., J. A. Joines, S. D. Roberts, K. A. Thoney, J. R. Wilson. 2000. Satisfying due-dates in large job shops: Characteristics of "real" problems. Preprint.

Iman, R. and W. Conover. 1982. A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics: Simulation and Computation*, 11: 311-334.

Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis, 3rd ed.* McGraw Hill, Boston.

Li, S. T., and J. L. Hammond. 1975. Generation of pseudo-random numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics.* 5:557–561.

Lurie, P. M., and M. S. Goldberg. 1998. An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science*. 44:203–218.

Mackenzie, G. R. 1994. *Approximately Maximum-Entropy Multivariate Distributions with Specified Marginals and Pairwise Correlations*. Ph.D. thesis. Department of Decision Sciences, University of Oregon, Eugene OR.

Mardia, K. V. 1970. A translation family of bivariate distributions and Fréchet's bounds. *Sankhya.* A32:119–122.

Meeuwissen, A. M. H., and T. Bedford. 1997. Minimally informative distributions with given rank correlation for use in uncertainty analysis. *Journal of Statistical Computation and Simulation* 57: 143–174.

Meeuwissen, A. M. H., and R. M. Cooke. 1994. Tree dependent random variables. Technical report 94-28, Department of Mathematics, Delft University of Technology, Delft, The Netherlands.

Walker, A. J. 1977. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software* 3: 253–256.

Whitt,W. 1976. Bivariate distributions with given marginals. *Annals of Statistics*. 4:1280–1289.

AUTHOR BIOGRAPHIES

SOUMYADIP GHOSH is a doctoral student in the School of Operations Research and Industrial Engineering at Cornell University.

SHANE G. HENDERSON is an assistant professor in the School of Operations Research and Industrial Engineering at Cornell University. He has previously held positions in the Department of Industrial and Operations Engineering at the University of Michigan and the Department of Engineering Science at the University of Auckland. He is an associate editor for the ACM Transactions on Modeling and Computer Simulation, and the assistant newsletter editor for the INFORMS College on Simulation. His research interests include discrete-event simulation, queueing theory and scheduling problems. His e-mail address is <shane@orie.cornell.edu>, and his web page can be found at <www.orie.cornell.edu/~shane>.